Causal inference with missing values in the covariates

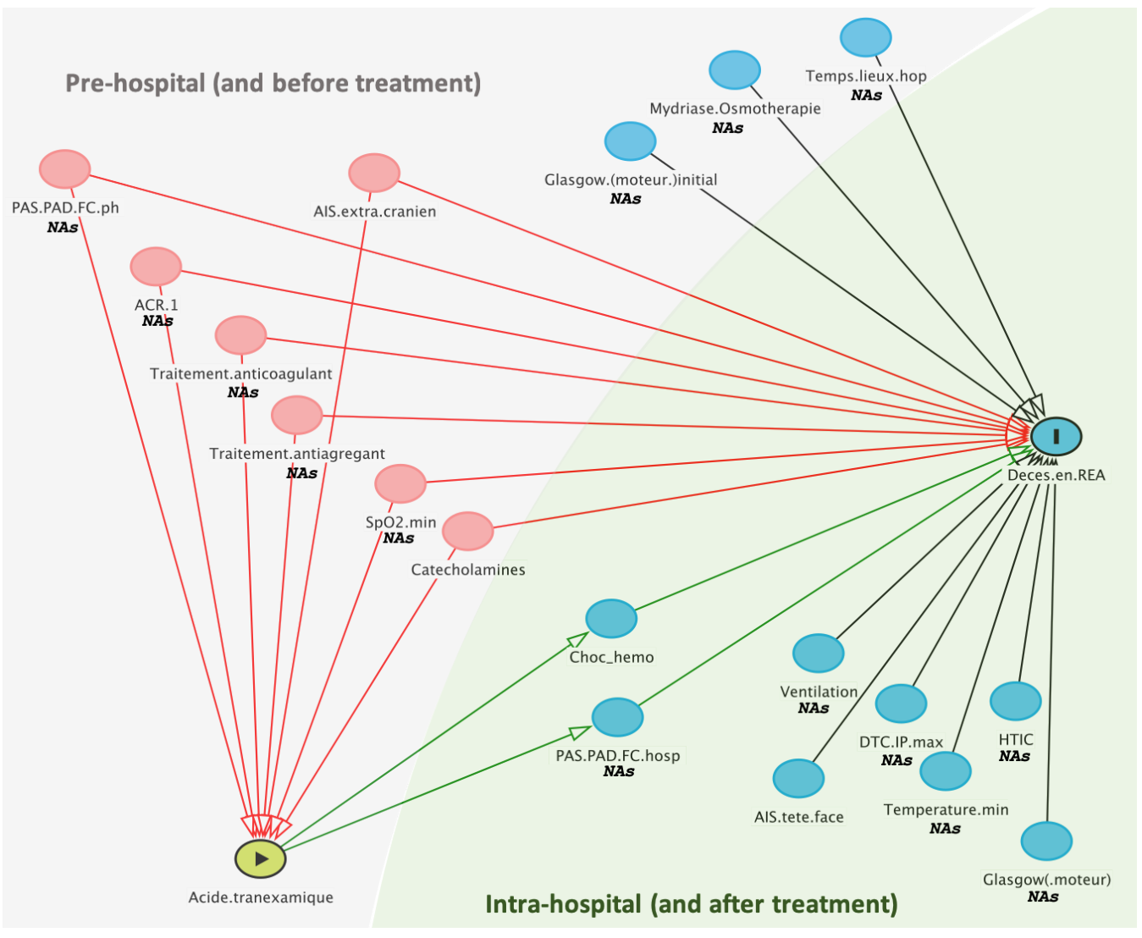
Treatment effect estimation of tranexamic acid on mortality for traumatic brain injury patients

IMKE MAYER(1,2), JULIE JOSSE(2), JEAN-PIERRE NADAL(1), TOBIAS GAUSS(3), JEAN-DENIS MOYER(3)

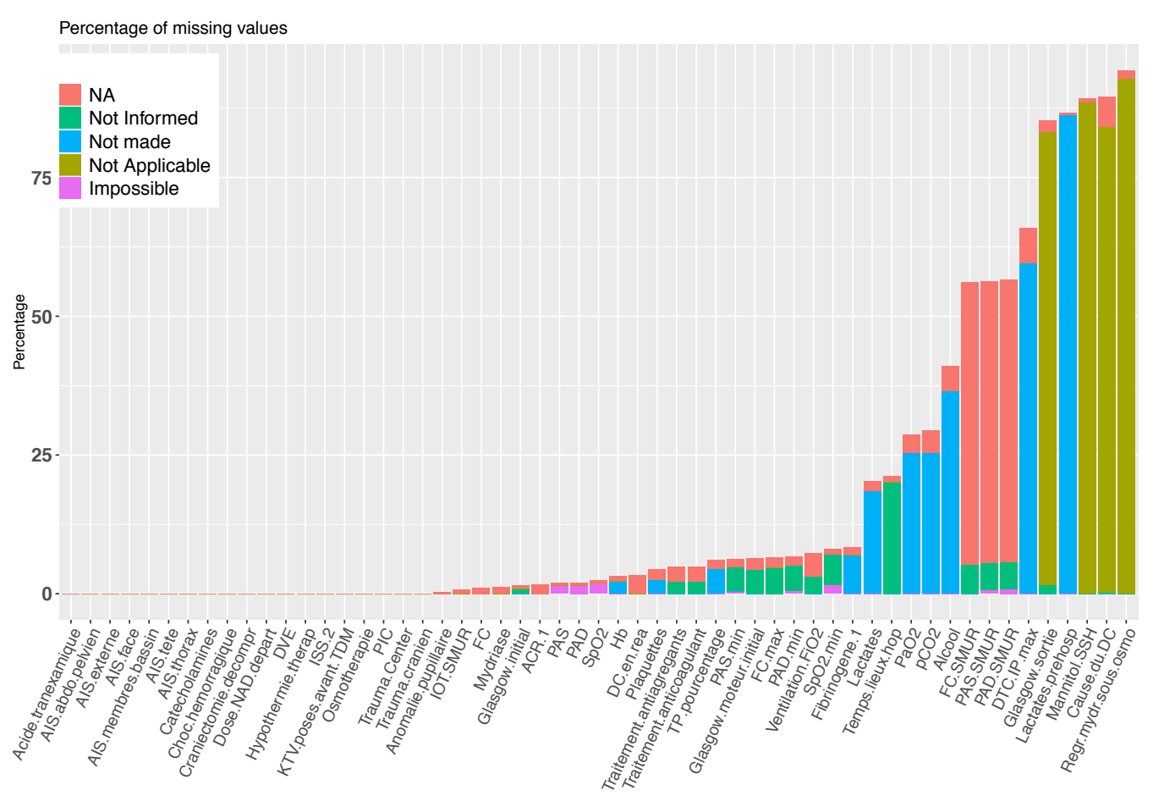
(1)École des Hautes Études en Sciences Sociales, (2)Ecole Polytechnique, (3) Traumabase® Group

MOTIVATIONS

Estimate the effect of tranexamic acid (TA) on the in-ICU mortality among patients with traumatic brain injury (TBI), based on the observational database **Traumabase®**. This database includes 7,945 major trauma patients, of which 3,050 have traumatic brain injury, with 244 pre-hospital and hospital measurements. The data is **heterogeneous**, being composed of both quantitative or categorical variables. Major trauma is a public health challenge and a major source of mortality and handicap around the world.


Pre-hospital (and before treatment)
Intra-hospital (and after treatment)
Acide tranexamique
Deces en.REA

Treatment effect (TE) estimation on **observational data** is challenging when the data contains **missing values**.

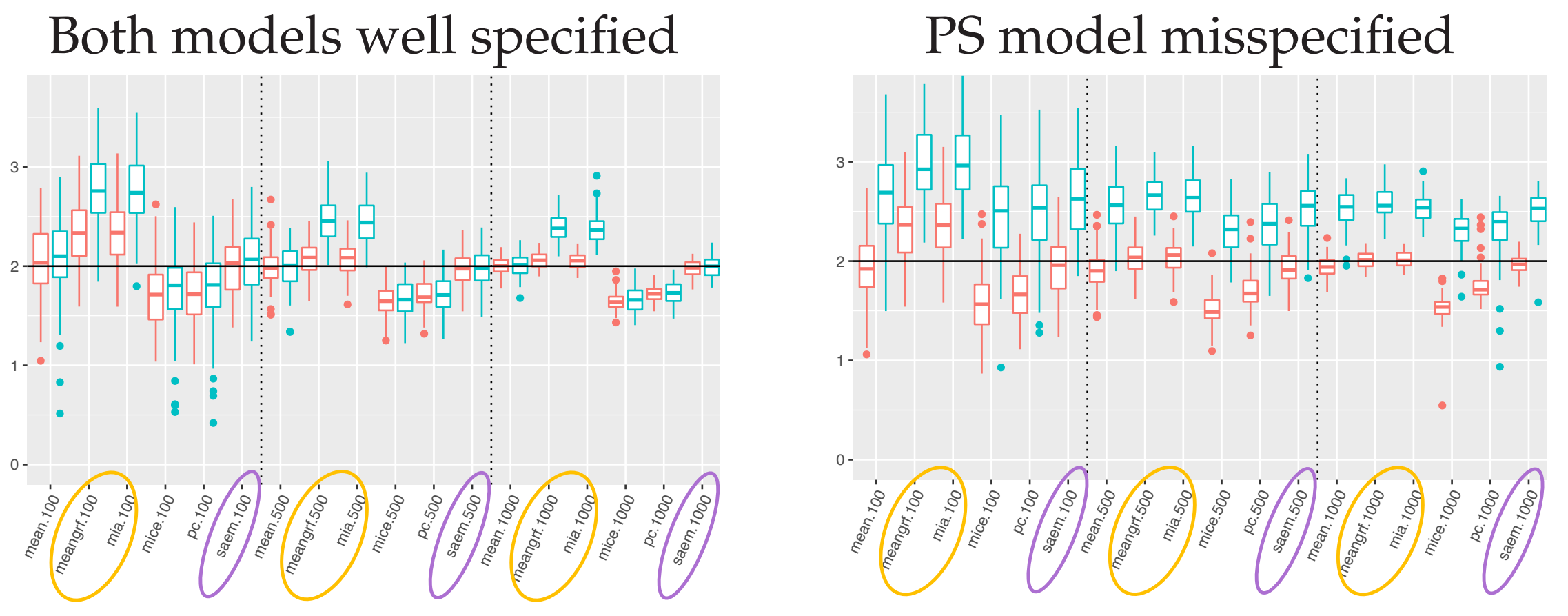


PROPOSAL

- Comparison of different TE estimators when covariates are partially observed, analysis of the bias.
- Proposition of new double robust TE estimator, based on random forests, handling missing values in the covariates.
- Application to critical care patient data.

SIMULATIONS

- $\mathbf{X} \sim \mathcal{N}(0,\Sigma)$ (s.t. $\Sigma_{ij} = \mathbb{1}_{i=j} + \rho \mathbb{1}_{i\neq j}$ with $\rho = 0.6$).
- Logistic-linear model for $T \in \{0,1\}, Y \in \mathbb{R}$, satisfying CIT.
- MAR (NA in X_1, X_3 depend on X_2).
- True ATE: $\tau = 2$.
- IPW and DR estimators.
- (Generalized) propensity score (PS) estimation with missing values:
  - (a) imputation (**mean**, **mice**, **principal component**) + logistic regression,
  - (b) logistic regression handling NAs (**SAEM**) [2],
  - (c) random forest with missing incorporated in attributes (**MIA**) or mean imputation.


Both models well specified

PS model misspecified

FUTURE RESEARCH

- Prove consistency / double robustness of the proposed ATE estimator handling missing values in the covariates (and heterogeneous data).
- TBI is very heterogeneous in terms of clinical presentation, pathophysiology and outcome → heterogeneous TE estimation.
- Long-term objective: developing a decision support tool for clinical care management.
- Compare results to the soon to be published randomized controlled trial *CRASH-3* results [1].

CAUSAL INFERENCE WITH MISSING VALUES IN THE COVARIATES

Assumptions:

→ **Rubin's potential outcome framework**: T binary treatment, $(Y_i(t))_{t\in\{0,1\}}$ potential outcomes.

$$\tau = \mathbb{E}[\Delta_i] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \quad \textbf{(ATE)},$$

$\mathbf{X} = (\mathbf{X}^{obs}, \mathbf{X}^{mis}) \in \mathbb{R}^{n\times p}$ completely observed confounders, $e(x) = \mathbb{P}(T=1\,|\,X=x)$ propensity score, $\mu_t(x) = \mathbb{E}[Y(t)|X=x]$ conditional response surface.

→ **Missing values**: $\mathbf{R} \in \{0,1\}^{n\times p}$ response indicator matrix, $\tilde{\mathbf{X}} = \mathbf{X}\odot\mathbf{R} + \texttt{NA}(1-\mathbf{R}) \in (\mathbb{R}\cup\texttt{NA})^{n\times p}$ observed confounders, $e^*(x,r) = \mathbb{P}(T=1\,|\,X^{obs}=x, R=r)$ generalized propensity score [7].

→ Classical causal inference assumptions: SUTVA, unconfoundedness, overlap.

→ Additional assumptions due to missingness:

- unconfoundedness*: $Y_i(t) \perp\!\!\!\perp T_i \,|\, X_i, R_i \quad t\in\{0,1\}$
- CIT or CIO: $T_i \perp\!\!\!\perp X_i^{mis} \,|\, X_i^{obs}, R_i$ or $Y_i(t) \perp\!\!\!\perp X_i^{mis} \,|\, X_i^{obs}, R_i \quad t\in\{0,1\}$

Method

→ Treatment effect estimator $\hat{\tau}_{DR,*}$ with **double robustness property** [6] (conjecture):

$$\hat{\tau}_{DR,*} = \frac{1}{n}\left(\sum_{i=1}^{n} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + T_i\frac{Y_i - \hat{\mu}_1(X_i)}{\hat{e}^*(X_i)} - (1-T_i)\frac{Y_i - \hat{\mu}_0(X_i)}{1-\hat{e}^*(X_i)}\right)$$

Propensity model (e^*) correctly specified:

$$\mathbb{E}\left[1 - \frac{T_i}{e^*(X_i)}\,|X_i^{obs}, R_i\right] = 0$$

$\Rightarrow \hat{\tau}_{DR,*} = \hat{\tau}_{IPW,*}$ is consistent.

Outcome model (μ) correctly specified:

$$\mathbb{E}\left[Y_i - \mu_1(X_i)|\; T_i = 1, X_i^{obs}, R_i\right] = 0$$

$\Rightarrow \hat{\tau}_{DR,*}$ is consistent.

→ Parametric or nonparametric estimation of $\mu_t(\cdot)$ and $e(\cdot)$ → **interpretability of $\hat{\tau}_{\mathrm{DR}}$** is the same.

→ Nonparametric estimation using **random forests** to handle heterogeneous data and missing values consistently [3].

FIRST RESULTS

On Traumabase:

→ 11 identified confounders (continuous & discrete & categorical).

→ 12% treated patients.

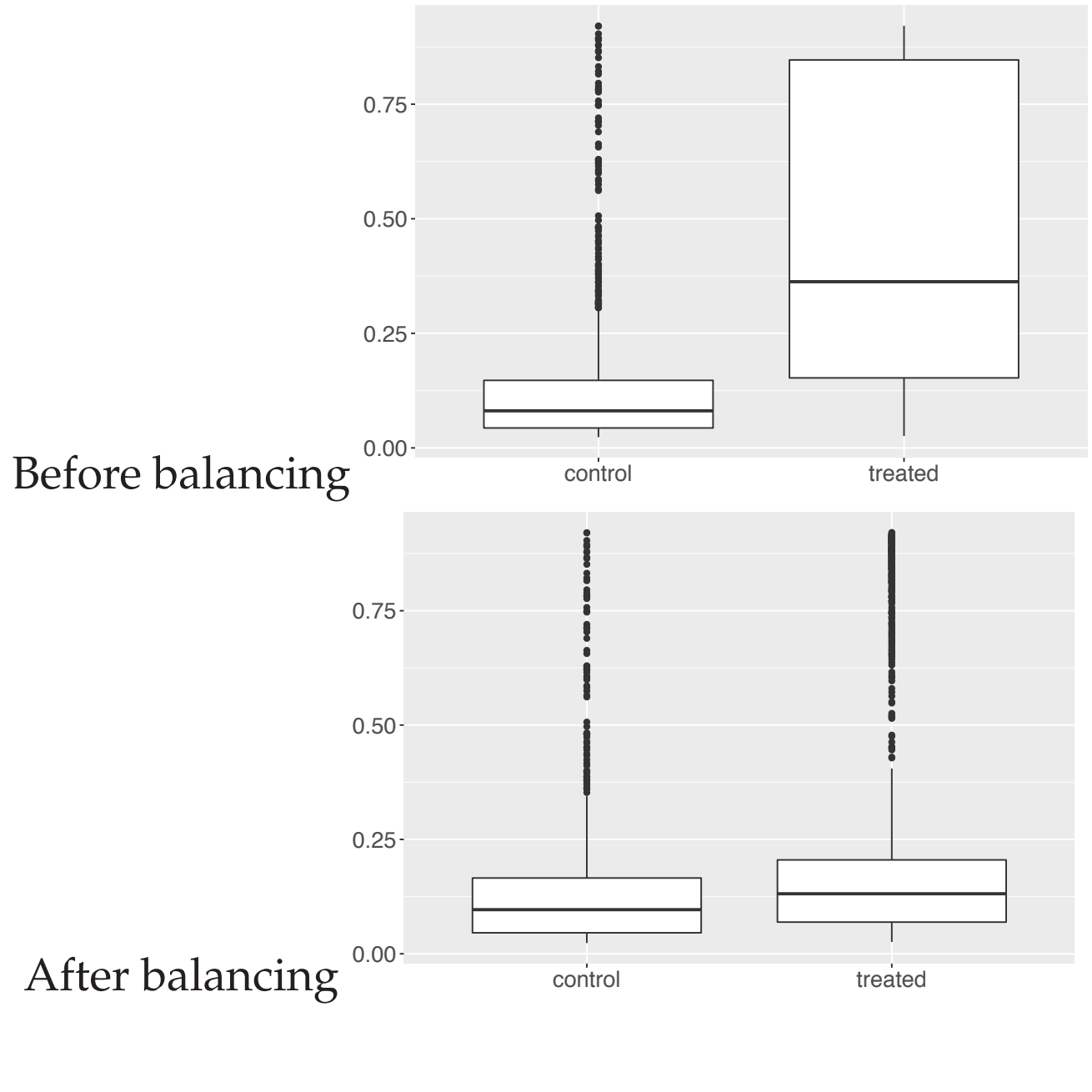
→ 0% - 23% of missing values (in confounders).

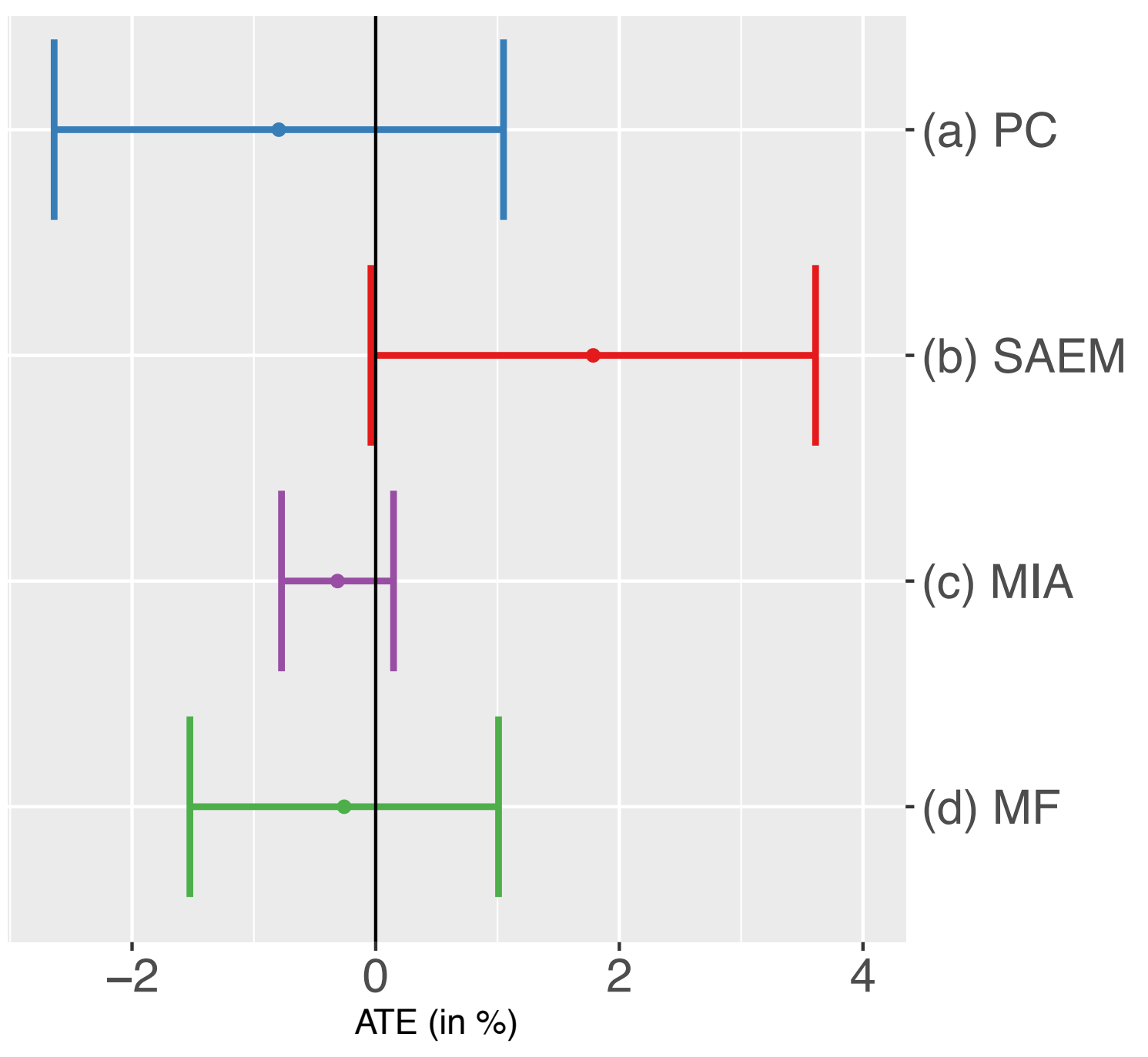
→ Different PS estimation techniques (logistic regression, gradient boosting, random forest).

→ 4 estimation approaches:
(a) Imputation (pc) + PS estimation
(b) PS estimation on incomplete data (SAEM)
(c) PS estimation via random forest with MIA
(d) Low-rank approximation + PS estimation [4]

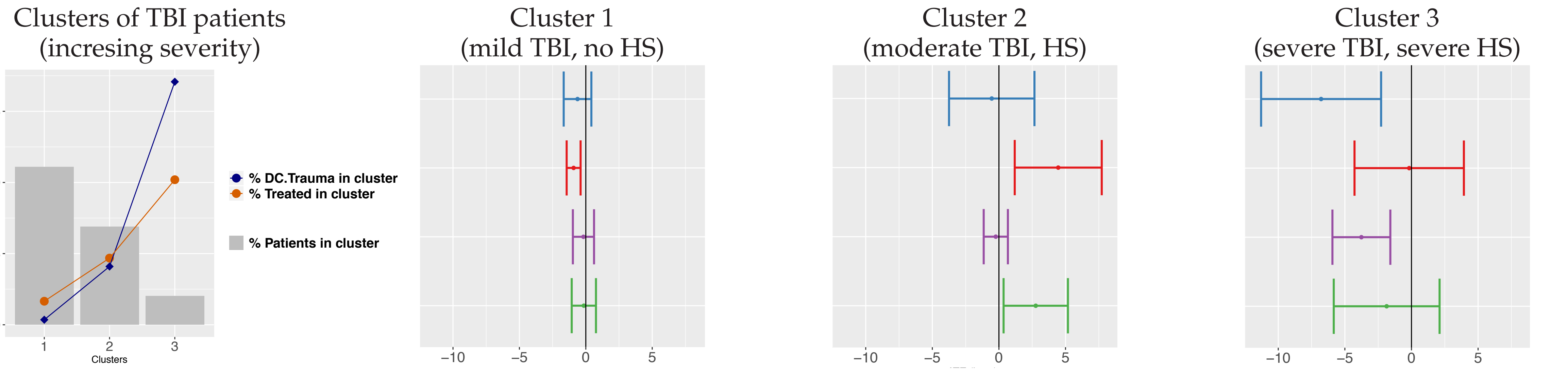
→ Handle overlap issues with overlap weights [5].

→ Identify patient clusters and estimate ATE.





- Difference in percentage points between mortality rates in treatment and control groups.
- **No evidence for rejecting null hypothesis of no effect of TA on in-ICU mortality among TBI patients.**
- Different TE w.r.t. severity of TBI and extra-cranial lesions.


Clusters of TBI patients (incresing severity)

Cluster 1 (mild TBI, no HS)

Cluster 2 (moderate TBI, HS)

Cluster 3 (severe TBI, severe HS)

REFERENCES

[1] Y. Dewan, E. O. Komolafe, J. H. Mejía-Mantilla, P. Perel, I. Roberts, and H. Shakur. Crash-3-tranexamic acid for the treatment of significant traumatic brain injury: study protocol for an international randomized, double-blind, placebo-controlled trial. *Trials*, 13(1):87, 2012.

[2] W. Jiang, J. Josse, and M. Lavielle. Logistic regression with missing covariates–parameter estimation, model selection and prediction. *arXiv preprint*, 2018.

[3] J. Josse, N. Prost, E. Scornet, and G. Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint*, 2019.

[4] N. Kallus, X. Mao, and M. Udell. Causal inference with noisy and missing covariats via matrix factorization. *arXiv preprint*, 2018.

[5] F. Li, K. L. Morgan, and A. M. Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

[6] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[7] P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

See also **R-miss-tastic**, a unified platform on missing values methods and workflows, `https://rmisstastic.netlify.com`.